Improved Converses and Gap-Results for Coded Caching
Chien-Yi Wang, Shirin Saeedi Bidokhti, and Michèle Wigger

Abstract
Improved lower bounds on the average and the worst-case rate-memory tradeoffs for the Maddah-Ali&Niesen coded caching scenario are presented. For any number of users and files and for arbitrary cache sizes, the multiplicative gap between the exact rate-memory tradeoff and the new lower bound is less than 2.315 in the worst-case scenario and less than 2.507 in the average-case scenario.

I. INTRODUCTION
We consider the canonical coded caching scenario by Maddah-Ali and Niesen [1] with a single transmitter and K receivers, where each receiver is equipped with a cache memory of equal size (see in Figure 1). Communication takes place in two phases. In a first caching phase the transmitter stores contents (arbitrary functions of files) at the receivers' cache memories. In a subsequent delivery phase it conveys one of the files stored in its library to each of the receivers. The main challenge in this configuration is that during the caching phase it is not known which receiver demands which specific file from the library. The cache contents thus need to be designed so as to be useful for many possible demands.

Traditional caching systems store the same most popular files in each and every cache memory. This allows the receivers to retrieve these files locally without burdening the common communication link from the transmitter to the receivers. Recently [1], it has been shown that much larger gains, so called global caching gains, are possible if the various receivers store different parts of the files in their cache memories. In this case, the transmitter can simultaneously serve multiple receivers during the delivery phase by sending coded data, and thus significantly reduced the required communication rate on the common link connecting the transmitter to the receivers.

The main quantity of interest in this work is the rate-memory tradeoff—i.e., the minimum required delivery rate, as a function of the available cache memories, so that all receivers reliably recover their demanded files. Several recent works [1]–[17] have presented upper and lower bounds on the rate-memory tradeoff. More specifically, the works in [1]–[12], [14] investigated the worst-case rate-memory tradeoff where the communication rate cannot depend on the specific demand. (The works in [4], [5], [14] determined the exact rate-memory tradeoff if the caching contents are restricted to parts of messages. This caching strategy is however known to be suboptimal in general.) The currently best lower and upper bounds for this worst-case rate-memory tradeoff have been shown to match up to a multiplicative gap of 4 [10].

The works in [13], [14] investigated the average rate-memory tradeoff where the communication rate can depend on the specific demand, and the rate of interest is the average required rate over all possible demands. The currently best lower and upper bounds for this average rate-memory tradeoff have been shown to match up to a multiplicative gap of $4.7$ [13].

In this paper we provide new lower bounds on the worst-case and the average rate-memory tradeoffs. The new lower bounds match the exact worst-case and average rate-memory tradeoffs up to multiplicative gaps of 2.315 and 2.507, respectively. More precisely, these gaps are with respect to the upper bounds on the rate-memory tradeoffs under decentralized caching in [14]. An upper bound on the rate-memory tradeoff under decentralized caching is also an upper bound on the rate-memory tradeoff under centralized caching considered here, because in the decentralized caching the cache content at a given receiver has to be chosen according to a specific distribution, whereas in centralized caching any content can be cached that satisfies the cache memory constraints.
Fig. 1: Coded caching scenario with $K$ receivers having equal cache size $FM$ bits.

Many other variations of the caching problem have been studied. For example, the online caching problem [18]; caching with non-uniform demands [19]–[21]; caching of correlated files [22]–[24] where Wyner’s and Gac-Körner’s common information play an important role [22]; caching in noisy broadcast channels [25]–[27], [29]–[43] where performance can be improved through joint cache-channel coding [25], [26], [32], state-information [33], [34], [34], [37], [39], or massive MIMO [36]; caching in Gaussian interference networks [44]–[46]; caching in hierarchical networks [54], [56], [57]; and caching in cellular networks [46]–[53] where it allows to cancel inter-cell interference [47].

II. DETAILED PROBLEM SETUP

The transmitter has a library of $N$ independent messages $W_1, \ldots, W_N$, where each $W_d$ is uniformly distributed over $\{1, \ldots, 2^F\}$ for $F$ a positive integer. Each of the $K$ receivers is provided with a cache memory of size $FM$ bits, see Figure 1. Suppose that each receiver will demand exactly one message from the library. We denote the demand of receiver $k$ by

$$d_k \in \mathcal{N} := \{1, \ldots, N\},$$

and thus the message demanded by receiver $k$ is $W_{d_k}$. Let

$$\mathbf{d} := (d_1, \ldots, d_K)$$

denote the receivers’ demand vector. The communications process takes place in two phases: a caching phase and a delivery phase.

During a period of low network-congestion and before the receivers’ demand vector $\mathbf{d}$ is known, the transmitter sends an individual cache message $\forall_k \in \{1, \ldots, [2^FM]\}$, to each of the $K$ receivers. Since $\mathbf{d}$ is unknown at this time, the cache messages will be functions of the entire library. For every $k \in \{1, \ldots, K\}$:

$$\forall_k := g_k(W_1, \ldots, W_N),
$$

$$g_k : \{1, \ldots, 2^F\}^N \to \{1, \ldots, [2^FM]\}.$$  

In the delivery phase, the transmitter is given the receivers’ demands $\mathbf{d} = (d_1, \ldots, d_K)$, and it generates the delivery-symbol $X$ sent over the common noise-free bit-pipe as:

$$X := f_d(W_1, \ldots, W_N),
$$

$$f_d : \{1, \ldots, 2^F\}^N \to \mathcal{X},$$

where $\mathcal{X}$ is the delivery alphabet that we will specify shortly.

We also assume that $\mathbf{d}$ is known to all receivers (e.g., $\mathbf{d}$ can be communicated to the receivers with zero transmission rate). Each receiver $k \in \{1, \ldots, K\}$ perfectly observes the delivery-symbol $X$, and can thus recover its desired message as

$$\hat{W}_k := \varphi_{k,\mathbf{d}}(X, \forall_k),$$

\textsuperscript{1}Allowing for randomized caching functions does not change the results of this paper.
\[ \phi_{k,d} : \mathcal{X} \times \{1, \ldots, |2^{FM}|\} \rightarrow \{1, \ldots, 2^F\}. \]

We are left with specifying the delivery alphabet \( \mathcal{X} \). We distinguish two scenarios:

- In the worst-case scenario, the delivery alphabet \( \mathcal{X} \) does not depend on the demand vector \( d \).
  - In this scenario, a rate-memory pair \((R, M)\) is achievable if for every \( \epsilon > 0 \) and sufficiently large message lengths \( F \), there exist caching, encoding, and decoding functions with delivery alphabet
    \[ \mathcal{X} = \{1, \ldots, |2^{F(R + \epsilon)}|\}, \tag{1} \]
    so that for each demand vector \( d \in \mathcal{N}^K \), each receiver \( k \) can perfectly reconstruct its desired message:
    \[ \hat{W}_k = W_{d_k}. \tag{2} \]

- In the average-case scenario, the delivery alphabet \( \mathcal{X} \) depends on the demand vector \( d \).
  - In this scenario, a rate-memory pair \((R, M)\) is achievable if for each demand vector \( d \in \mathcal{N}^K \), each \( \epsilon > 0 \), and sufficiently large message lengths \( F \), there exist caching, encoding, and decoding functions with delivery alphabet
    \[ \mathcal{X}_d = \{1, \ldots, |2^{FR_d}|\}, \tag{3} \]
    so that each Receiver \( k \in \{1, \ldots, K\} \) can perfectly reconstruct its desired message, (2), and
    \[ \frac{1}{N^K} \sum_{d \in \mathcal{N}^K} R_d \leq R + \epsilon. \tag{4} \]

The main focus of this paper is on the rate-memory tradeoffs of the worst-case and the average-case scenarios:

**Definition 1:** Given the cache memory size \( M \), we define the rate-memory tradeoffs \( R^*_{\text{worst}}(M) \) and \( R^*_{\text{avg}}(M) \) as the infimum of all rates \( R \) such that the rate-memory pair \((R, M)\) is achievable for the worst-case and the average-case scenarios, respectively.

The problem is interesting for \( M < N \).

For \( M \geq N \) each receiver can store all the library in its cache memory and there is no need to transmit anything during the delivery phase. We henceforth assume that (5) holds.

### III. Main Results

Define \( \bar{N} := \min\{K, N\} \) and \( \bar{N} := \{1, 2, \ldots, \bar{N}\} \).

#### A. Worst-Case Scenario

Our first result is a lower bound on the rate-memory tradeoff in the worst-case scenario. It is proved in Section IV-B. Alternatively, it can also be extracted from the converse result for general degraded broadcast channels in [28].

**Theorem 1:** For all \( M \in [0, N] \),
\[ R^*_{\text{worst}}(M) \geq R^\text{low}_{\text{worst}}(M), \tag{6} \]
where
\[
R^\text{low}_{\text{worst}}(M) := \max \left\{ \max_{\ell \in \bar{N}} \left[ \ell - M \frac{\ell^2}{N} \right], \right. \\
\left. \max_{\ell \in \bar{N}} \left[ \ell - M \sum_{k=1}^{\ell} \frac{k}{N-k+1} \right] \right\}.
\]

Figure 2 compares this new lower bound on \( R^*_{\text{worst}}(M) \) with the existing lower bounds in [1] and [6]. The figure also shows upper bounds from [14]. The red solid upper bound is for centralized caching, as considered in this paper. The green dashed upper bound is for decentralized caching. For simplicity, the latter upper bound is used to derive the gap results in the following Theorem 2.
Fig. 2: Upper and lower bounds on $R_{\text{worst}}^*(M)$ for $K = 16$ and $N = 64$.

**Theorem 2:** Irrespective of the number of users $K$, the library size $N$, and the memory size $M \in [0, N)$:

$$\frac{R_{\text{worst}}^*(M)}{R_{\text{low}}^*(M)} \leq \max_{\ell \in \mathbb{Z}^+} \max_{a \in (0,1)} \phi(a, \ell),$$

where

$$\phi(a, \ell) := \frac{a(\ell+1)}{(1-a)^{\ell}} \left( 1 - \left( \frac{\ell+a}{\ell+1} \right)^{\ell/a} \right) \leq 2.315$$

**Proof:** See Section V. $

\blacksquare$

**Remark 3:** For any $\ell \in \mathbb{Z}^+$, the function $a \mapsto \phi(a, \ell)$ is continuous and bounded over $(0,1)$, see also Figure 3. Numerical evaluation\(^2\) shows that for $\ell \in \{1, \ldots, 10^4\}$:

$$\max_{a \in (0,1)} \phi(a, \ell) \leq 2.315.$$  

Moreover,

$$\max_{\ell \in \mathbb{Z}^+: \ell > 10^4} \max_{a \in (0,1)} \phi(a, \ell) \leq \max_{b \in (0,10^{-4})} \max_{a \in (0,1)} \psi(a, b),$$

where

$$\psi(a, b) := \frac{a(1+b)}{1-a} \left( 1 - \left( \frac{1+ab}{1+b} \right)^{\frac{1}{ab}} \right) \left( 1 - \frac{(1-a)b}{1-a+ab} + \frac{1-a}{a} \ln \left( 1 - a + ab \right) \right).$$

The function $(a, b) \mapsto \psi(a, b)$ is continuous and bounded over $(0,1) \times (0, 10^{-4})$, see also Figure 4. Numerical evaluation shows that

$$\max_{b \in (0,10^{-4})} \max_{a \in (0,1)} \psi(a, b) \leq 2.315.$$ \hspace{1cm} (9)

**Proof:** Inequality (7) is proved in Section V-B. $

\blacksquare$

\(^2\)All numerical evaluations in this paper are performed by applying the MATLAB function `fmincon` with the sequential quadratic programming (SQP) method.
Fig. 3: The functions $\phi(a, \ell)$ for $a \in (0, 1)$ and $\ell = 1, \ldots, 10^4$.

Fig. 4: The function $\psi(a, b)$ over $(a, b) \in (0, 1) \times (0, 10^{-4})$.

B. Average-Case Scenario

Theorem 4: For all $M \in [0, N)$,

$$R^*_\text{avg}(M) \geq R^\text{low}_\text{avg}(M),$$

where

$$R^\text{low}_\text{avg}(M) := \max \left\{ \max_{\ell \in \{1, \ldots, K\}} \left[ \left(1 - \left(1 - \frac{1}{N}\right)^\ell\right)(N - \ell M) \right], \right.$$

$$\left. \max_{\ell \in \{1, \ldots, K\}} \left[ \left(1 - \left(1 - \frac{1}{N}\right)^\ell\right)N - \frac{\ell(\ell + 1)}{2N} M \right] \right\}.$$

Proof: See Section IV-C.

Figure 5 compares this new lower bound on $R^*_\text{avg}(M)$ with the existing lower bounds in [13] and the upper bounds in [14]. The red solid upper bound is for centralized caching, as considered in this paper. The green dashed upper bound is for decentralized caching and also from [14].
As the following theorem and remark show, the multiplicative gap between the lower bound of Theorem 4 and $R^*_{\text{avg}}(M)$ is at most 2.507.

**Theorem 5:** Irrespective of the number of users $K$, the library size $N$, and the memory size $M \in [0, N)$:

$$\frac{R^*_{\text{avg}}(M)}{R^\text{low}_{\text{avg}}(M)} \leq \max_{u \in (0,1]} \max_{v \in (0,1/2]} \eta(u, v),$$

where

$$\eta(u, v) := \frac{(u + v - v(1-v)^{\frac{u}{u+v}}) \left(1 - \left(1 - \frac{v}{u+v} (1-v)^{\frac{u}{u+v}}\right)^{\frac{1}{v}}\right)}{(1-v)^{\frac{u}{v}} \left(1 - (1 + \frac{u}{v}) (1-v)^{\frac{u}{v}}\right)}.$$

**Proof:** See Section VI. \hfill \blacksquare

**Remark 6:** The function $\eta$ is continuous and bounded over $(0,1] \times (0,1/2]$. Numerical evaluation shows that

$$\max_{u \in (0,1]} \max_{v \in (0,1/2]} \eta(u, v) \leq 2.507. \quad (10)$$

**IV. PROOF OF THEOREMS 1 AND 4**

A. Auxiliary Lemmas

The following two lemmas will be used in the proofs of Theorems 1 and 4.

The next lemma is stated for the average-case scenario. It applies readily also to the worst-case scenario if rate $R_d$ is replaced by $R$.

**Lemma 7:** Fix a number $\ell \in \mathcal{N}$ and a demand vector $d \in \mathcal{N}^K$ whose first $\ell$ entries are $d_1, \ldots, d_\ell$. Fix also a small $\epsilon > 0$ and assume a sufficiently large $F$ with caching, encoding, and decoding functions so that (2) holds for all $k \in \{1, \ldots, K\}$. Then,

$$R_d + \epsilon \geq \kappa_d(\ell) - \frac{1}{F} \sum_{k=1}^{\ell} I(W_{d_k}; V_1, \ldots, V_k|W_{d_1}, \ldots, W_{d_{k-1}}), \quad (11)$$

where $\kappa_d(\ell)$ denotes the number of distinct demands for receivers $1, \ldots, \ell$:

$$\kappa_d(\ell) := |\{d_1, \ldots, d_\ell\}|. \quad (12)$$
Fig. 6: The function \( \eta(u,v) \) over \( (u,v) \in (0,1) \times (0,1) \).

**Proof:** For any \( k \in \{1,\ldots,\ell\} \):

\[
I(X; W_{d_k} | \forall_1, \ldots, \forall_k, W_{d_1}, \ldots, W_{d_{k-1}})
\]

\( \overset{(a)}{=} \) \( H(W_{d_k} | \forall_1, \ldots, \forall_k, W_{d_1}, \ldots, W_{d_{k-1}}) \)

\( = H(W_{d_k} | W_{d_1}, \ldots, W_{d_{k-1}}) \)

\( - I(W_{d_k}; \forall_1, \ldots, \forall_k | W_{d_1}, \ldots, W_{d_{k-1}}) \)

\( \overset{(b)}{=} F \cdot 1 \{ d_k \notin \{ d_1, \ldots, d_{k-1} \} \}

- I(W_{d_k}; \forall_1, \ldots, \forall_k | W_{d_1}, \ldots, W_{d_{k-1}}), \)

(13)

where (a) holds because (2) implies that \( H(W_{d_k} | X, \forall_1, \ldots, \forall_k, W_{d_1}, \ldots, W_{d_{k-1}}) = 0 \); and (b) holds by the independence of the messages and because \( H(W_d) = F \) for any \( d \in \mathcal{N} \).

On the other hand,

\[
\sum_{k=1}^{\ell} I(X; W_{d_k} | \forall_1, \ldots, \forall_k, W_{d_1}, \ldots, W_{d_{k-1}})
\]

\( \leq \sum_{k=1}^{\ell} I(X; W_{d_k}, \forall_k | \forall_1, \ldots, \forall_{k-1}, W_{d_1}, \ldots, W_{d_{k-1}}) \)

\( = I(X; W_{d_1}, \ldots, W_{d_\ell}, \forall_1, \ldots, \forall_\ell) \)

\( \leq H(X) \leq F(R_d + \epsilon). \)

(14)

Combining (13) and (14) establishes the lemma.

**Lemma 8:** Let \( L \) be a positive integer, \( A_1, \ldots, A_L \) be an independent random \( L \)-tuple, and \( \forall \) be a random variable arbitrarily correlated with \( A_1, \ldots, A_L \). For any subset \( S \subseteq \{1,\ldots,L\} \), denote by \( A_S \) the subset \( \{ A_s, s \in S \} \). Then, for all \( l \in \{1,\ldots,L\} \),

\[
\frac{1}{\binom{L}{l}} \sum_{\substack{S \subseteq \{1,\ldots,L\} : |S| = l}} I(A_S; \forall) \leq \frac{l}{L} I(A_1, \ldots, A_L; \forall). \]

(15)

**Proof:** Consider any \( l \in \{1,\ldots,L\} \). We have

\[
\frac{1}{\binom{L}{l}} \sum_{\substack{S \subseteq \{1,\ldots,L\} : |S| = l}} I(A_S; \forall)
\]


\( \frac{1}{L} \sum_{S \subseteq \{1, \ldots, L\}: |S| = l} \sum_{j \in S} H(A_j) \)

\( \leq \frac{1}{L} \sum_{S \subseteq \{1, \ldots, L\}: |S| = l} H(A_S|V) \)

\( = \frac{(L-1)}{L} \sum_{j=1}^L H(A_j) - \frac{1}{L} \sum_{S \subseteq \{1, \ldots, L\}: |S| = l} H(A_S|V) \)

\( \leq \frac{L}{L} \sum_{j=1}^L H(A_j) - \frac{L}{L} H(A_1, \ldots, A_L|V) \)

\( = \frac{L}{L} I(A_1, \ldots, A_L; V), \)

where (a) and (c) follow since \( A_1, \ldots, A_L \) are independent and (b) follows from the generalized Han Inequality (see [60, Theorem 17.6.1]).

\[ R + \epsilon \geq \sum_{k=1}^\ell \alpha_k, \]

(17)

where

\[ \alpha_1 := \frac{1}{(N_\ell)!} \sum_{d \in Q_\ell^{\text{dist}}} \frac{1}{F} I(W_{d_1}; V_1), \]

(18a)

and for \( k = 2, \ldots, \ell : \)

\[ \alpha_k := \frac{1}{(N_\ell)!} \sum_{d \in Q_\ell^{\text{dist}}} \frac{1}{F} I(W_{d_k}; V_1, \ldots, V_k|W_{d_1}, \ldots, W_{d_{k-1}}). \]

(18b)

The following lemma and letting \( \epsilon \to 0 \) and thus \( F \to \infty \), concludes the proof.

**Lemma 9:** Parameters \( \alpha_1, \ldots, \alpha_\ell \) satisfy

\[ \sum_{k=1}^\ell \alpha_k \leq \min \left\{ \frac{\ell^2}{N}, \sum_{k=1}^\ell \frac{kM}{N - k + 1} \right\}. \]

(19)

**Proof:** We first prove that for each \( k \in \{1, \ldots, \ell\} : \)

\[ \alpha_k \leq \frac{kM}{N - k + 1}, \]

(20)

The following lemma and letting \( \epsilon \to 0 \) and thus \( F \to \infty \), concludes the proof.

**Lemma 9:** Parameters \( \alpha_1, \ldots, \alpha_\ell \) satisfy

\[ \sum_{k=1}^\ell \alpha_k \leq \min \left\{ \frac{\ell^2}{N}, \sum_{k=1}^\ell \frac{kM}{N - k + 1} \right\}. \]

(19)

**Proof:** We first prove that for each \( k \in \{1, \ldots, \ell\} : \)

\[ \alpha_k \leq \frac{kM}{N - k + 1}, \]

(20)

\[ \text{In (11) } R_d \text{ needs to be replaced by } R \text{ because here we consider a worst-case scenario.} \]
which establishes the upper bound
\[
\sum_{k=1}^{\ell} \alpha_k \leq \sum_{k=1}^{\ell} \frac{kM}{N-k+1}.
\]  

For each partial demand vector \( \tilde{d} = (d_1, \ldots, d_{k-1}) \), let \( W_{\tilde{d}} := \{W_{d_1}, \ldots, W_{d_{k-1}}\} \). We have:
\[
F_{\alpha_k} = \frac{1}{\ell!(N/\ell)} \sum_{d \in Q_{\ell}^{\text{dist}}} \sum_{d \in Q_{\ell}^{\text{dist}}} I(W_{d_\ell}; \forall_1, \ldots, \forall_k|W_{d_1}, \ldots, W_{d_{k-1}})
\]
\[
= \frac{1}{\ell!(N/\ell)} \sum_{d \in Q_{\ell}^{\text{dist}}} \sum_{d \in Q_{\ell}^{\text{dist}}} I(W_{d_\ell}; \forall_1, \ldots, \forall_k|W_{\tilde{d}})
\]
\[
= \frac{1}{\ell!(N/\ell)} \sum_{d \in Q_{\ell}^{\text{dist}}} \sum_{d \in Q_{\ell}^{\text{dist}}} I(W_{d_\ell}; \forall_1, \ldots, \forall_k|W_{\tilde{d}})
\]
\[
= \frac{1}{\ell!(N/\ell)} \sum_{d \in Q_{\ell}^{\text{dist}}} \sum_{d \in Q_{\ell}^{\text{dist}}} \left[ H(W_1, \ldots, W_N|W_{\tilde{d}}) - \sum_{j \in N \setminus \tilde{d}} H(W_j|\forall_1, \ldots, \forall_k, W_{\tilde{d}}) \right]
\]
\[
\leq \frac{1}{\ell!(N/\ell)} \sum_{d \in Q_{\ell}^{\text{dist}}} I(W_1, \ldots, W_N; \forall_1, \ldots, \forall_k|W_{\tilde{d}})
\]
\[
\leq \frac{(k-1)!}{k!(N/\ell)} kFM
\]
\[
= \frac{kFM}{N-k+1},
\]  

where (a) holds because for each value of \( \ell \) and \( j \) there are \( \binom{N-k}{\ell-k}(\ell-k)! \) ordered demand vectors \( d = (d_1, \ldots, d_K) \in Q_{\ell}^{\text{dist}} \) with \( (d_1, \ldots, d_{k-1}) = \tilde{d} \) and with \( d_k = j \); (b) holds by the independence of the messages; (c) holds because for any random tuple \( (A_1, \ldots, A_L) \) it holds that \( \sum_{l=1}^{L} H(A_l) \geq H(A_1, \ldots, A_L) \); and (d) holds because \( I(W_1, \ldots, W_N; \forall_1, \ldots, \forall_k|W_{\tilde{d}}) \) cannot exceed \( kFM \). This concludes the proof of (20) and thus of (21).

We now prove
\[
\sum_{k=1}^{\ell} \alpha_k \leq \frac{\ell^2 M}{N}.
\]  

For each \( d \in Q_{\ell}^{\text{dist}} : \)
\[
I(W_{d_1}; \forall_1) + \sum_{k=2}^{\ell} I(W_{d_1}; \forall_1, \ldots, \forall_k|W_{d_1}, W_{d_2}, \ldots, W_{d_{k-1}})
\]
\[
\leq I(W_{d_1}, W_{d_2}, \ldots, W_{d_\ell}; \forall_1, \ldots, \forall_\ell).
\]  

So,
\[
F \left( \binom{N}{\ell} \right) \ell! \cdot \sum_{k=1}^{\ell} \alpha_k
\]
= \sum_{d \in Q_{\ell}^{rep}} I(W_{d_1}; V_1)
+ \sum_{k=2}^{\ell} I(W_{d_k}; V_1, \ldots, V_k|W_{d_1}, W_{d_2}, \ldots, W_{d_{k-1}})
\leq \sum_{d \in Q_{\ell}^{rep}} I(W_{d_1}, W_{d_2}, \ldots, W_{d_\ell}; V_1 \ldots, V_\ell)
= \ell! \sum_{d \in Q_{\ell}^{rep}, d_1 < d_2 \cdots < d_\ell} I(W_{d_1}, W_{d_2}, \ldots, W_{d_\ell}; V_1 \ldots, V_\ell)
\leq \ell! \left( \frac{N}{\ell} \right)^{\ell} I(W_1, \ldots, W_N; V_1, \ldots, V_\ell)
\leq \frac{\ell!}{N} \ell! \left( \frac{N}{\ell} \right)^{\ell} FM,

where (a) follows from Lemma 8.

\[ \text{C. Proof of Theorem 4} \]

For any $\ell \in \{1, \ldots, K\}$, let $Q_{\ell}^{rep}$ be the set of all ordered length-$\ell$ vectors $(d_1, \ldots, d_\ell) \in \mathcal{N}^\ell$, where repetitions are allowed. Notice that:

\[ |Q_{\ell}^{rep}| = N^\ell. \tag{25} \]

Recall also that in the average-case scenario under investigation, the demand vector $d := (d_1, \ldots, d_K)$ is uniform over $Q_{K}^{rep}$. Let $D := (D_1, \ldots, D_K) \sim \text{Uniform}(N^K)$.

Fix now an $\ell \in \{1, \ldots, K\}$, and average Inequality (11) over all demand vectors $d \in Q_{K}^{rep}$. This yields:

\[ \mathbb{E}_D[R_D + \epsilon] \geq \mathbb{E}_D[k_D(\ell)] - \sum_{k=1}^{\ell} \beta_k, \tag{26} \]

where

\[ \beta_1 := \frac{1}{F} I(W_{D_1}; V_1|D), \tag{27a} \]

and for $k = 2, \ldots, \ell$:

\[ \beta_k := \frac{1}{F} I(W_{D_k}; V_1, \ldots, V_k|W_{D_1}, \ldots, W_{D_{k-1}}, D). \tag{27b} \]

The following two lemmas and letting $\epsilon \to 0$ and thus $F \to \infty$, conclude the proof.

\[ \text{Lemma 10:} \]

\[ \mathbb{E}_D[k_D(\ell)] = N \left( 1 - \left( 1 - \frac{1}{N} \right)^{\ell} \right). \tag{28} \]

\[ \text{Proof:} \]

\[ \mathbb{E}_D[k_D(\ell)] = \mathbb{E}_D \left[ \sum_{k=1}^{\ell} \mathbbm{1}\{D_k \notin \{D_1, \ldots, D_{k-1}\}\} \right]
= \sum_{k=1}^{\ell} \mathbb{E}_D \left[ \mathbbm{1}\{D_k \notin \{D_1, \ldots, D_{k-1}\}\} \right]
= \sum_{k=1}^{\ell} (1 - \frac{1}{N})^{k-1}
= N \left( 1 - \left( 1 - \frac{1}{N} \right)^{\ell} \right). \tag{29} \]
Lemma 11: Parameters $\beta_1, \ldots, \beta_\ell$ satisfy
\[
\sum_{k=1}^\ell \beta_k \leq \min \left\{ \mathbb{E}_D[\kappa_D(\ell)] \cdot \frac{\ell M}{N}, \sum_{k=1}^\ell \frac{kM}{N} \right\}. \tag{30}
\]

Proof: We first prove that for each $k \in \{1, \ldots, \ell\}$:
\[
\beta_k \leq \frac{kM}{N}, \tag{31}
\]
which establishes the upper bound
\[
\sum_{k=1}^\ell \beta_k \leq \sum_{k=1}^\ell \frac{kM}{N}. \tag{32}
\]

Defining $D_k := (D_1, \ldots, D_k)$, we have:
\[
F \beta_k = I(W_{D_k}; V_1, \ldots, V_k|W_{D_1}, \ldots, W_{D_{k-1}}, D)
= I(W_{D_k}; V_1, \ldots, V_k|W_{D_1}, \ldots, W_{D_{k-1}}, D_k)
= \frac{1}{N^k} \sum_{d \in Q_{k}^{n_{1}}} \sum_{j=1}^{N} I(W_j; V_1, \ldots, V_k|W_d)
\leq \frac{1}{N^k} \sum_{d \in Q_{k}^{n_{1}}} kFM
= kFM,
\]
where (a) holds because the messages are independent and because $H(A_1, \ldots, A_L) \leq \sum_{i=1}^L H(A_i)$ for any random $L$-tuple $(A_1, \ldots, A_L)$; and (b) holds because $I(W_1, \ldots, W_N; V_1, \ldots, V_k|W_d)$ cannot exceed $kFM$. This concludes the proof of (31) and thus (32).

We now prove
\[
\sum_{k=1}^\ell \beta_k \leq \mathbb{E}_D[\kappa_D(\ell)] \cdot \frac{\ell M}{N}. \tag{33}
\]

Let $D^\text{dist}_\ell$ be a vector containing all distinct elements of $D_\ell := (D_1, \ldots, D_\ell)$. Notice that $D^\text{dist}_\ell$ is of length $\kappa_D(\ell)$. Also, following the definition of the previous section, $W_{D_\ell} := \{W_{D_1}, \ldots, W_{D_\ell}\} = W_{D^\text{dist}_\ell}$. We have:
\[
F \sum_{k=1}^\ell \beta_k
= I(W_{D_\ell}; V_1|D)
+ \sum_{k=2}^\ell \sum_{k=2}^\ell I(W_{D_k}; V_1, \ldots, V_k|W_{D_1}, W_{D_2}, \ldots, W_{D_{k-1}}, D)
\leq I(W_{D_\ell}; V_1, \ldots, V_\ell|D)
\leq I(W_{D_\ell}; V_1, \ldots, V_\ell|D_\ell)
= I(W_{D_\ell}; V_1, \ldots, V_\ell|D_\ell, \kappa_D(\ell))
\]
\[\sum_{i=1}^{\ell} P(\kappa D_i(\ell) = i) I(W_{D_i}; V_1, \ldots, V_\ell | D_\ell, \kappa D_i(\ell) = i) \]

\[= \sum_{i=1}^{\ell} P(\kappa D_i(\ell) = i) I(W_{D_i}; V_1, \ldots, V_\ell | D_\ell, \kappa D_i(\ell) = i) \]

\[= \sum_{i=1}^{\ell} P(\kappa D_i(\ell) = i) \sum_{\tilde{d} \in Q^\text{dist}_i} \frac{1}{\binom{N}{i}} I(W_{\tilde{d}}; V_1, \ldots, V_\ell) \]

\[\leq \sum_{i=1}^{\ell} P(\kappa D(\ell) = i) \frac{i}{N} I(W_1, \ldots, W_N; V_1, \ldots, V_\ell) \]

\[\leq \sum_{i=1}^{\ell} P(\kappa D(\ell) = i) \cdot i \cdot \frac{\ell FM}{N} \]

\[= \mathbb{E}_D [\kappa D(\ell)] \frac{\ell FM}{N}, \]

where (a) holds because of the Markov chain \(V_1, \ldots, V_\ell - (W_{D_i}^\text{dist}, \kappa D_i(\ell) - D_\ell); (b) holds because given \(\kappa D_i(\ell) = i\) the probability that \(D_\ell^\text{dist}\) equals a specific vector \(\tilde{d} \in Q^\text{dist}_i\) equals \(\binom{N}{i}^{-1}; (c) follows from Lemma 8; and (d) follows since \(I(W_1, \ldots, W_N; V_1, \ldots, V_\ell)\) cannot be larger than \(\ell FM\). \(\blacksquare\)

V. PROOF OF THE GAP-RESULTS IN THEOREM 2 AND REMARK 3

A. Proof of Theorem 2

We wish to uniformly bound the gap

\[\xi(K, N, M) := \frac{R^*_\text{worst}(M)}{R^\text{low}_\text{worst}(M)}, \]

irrespective of \(K, N \geq 1\) and \(M \in [0, N)\).

We recall the achievable rate-memory tradeoff from [14, Corollary 2]. For any pair of positive integers \(K, N \geq 1\), define

\[R_{\text{YMA}}(K, N, M) := \begin{cases} \frac{N}{M}, & \text{if } M = 0, \\ \frac{N-M}{N} \left(1 - \frac{M}{N} \right)^N, & \text{if } M \in (0, N). \end{cases} \]

Since \(R_{\text{YMA}}(K, N, M)\) upper bounds the rate-memory tradeoff under a decentralized caching assumption [14], it must also upper bound the rate-memory tradeoff under centralized caching as considered here. (In fact, decentralized caching imposes additional constraints on the caching functions \(g_k\) compared to our setup here.) Thus, for any number of users \(K\) and files \(N\):

\[R^\text{worst}(M) \leq R_{\text{YMA}}(K, N, M), \quad M \in [0, N). \]

We thus have

\[\xi(K, N, M) \leq \frac{R_{\text{YMA}}(K, N, M)}{R^\text{low}_\text{worst}(M)} \leq \frac{R_{\text{YMA}}(K, N, M)}{R^\text{worst}_\text{worst}(K, N, M)}, \]

where we defined

\[R^\text{worst}_\text{worst}(K, N, M) := \max_{\ell \in \mathbb{N}} \sum_{j=1}^{\ell} \left(1 - \frac{jM}{N-j+1}\right) \]

(37)
and where the second inequality holds because for all $K, N, M$:

$$P_{\text{worst}}^\text{low}(M) \geq \max_{\ell \in \mathcal{N}} \left[ \ell - \sum_{j=1}^{\ell} \frac{jM}{N-j+1} \right] = R_{\text{worst}}(K, N, M).$$ (38)

Define

$$M_\ell := \begin{cases} N - \ell & \text{if } \ell \in \{0, 1, \ldots, \bar{N} - 1\}, \\ 0 & \text{if } \ell = \bar{N}. \end{cases}$$

Note that

$$0 = M_\bar{N} < M_{\bar{N}-1} < \cdots < M_0 = N.$$ The function $R_{\text{worst}}(K, N, M)$ is piecewise-linear with $\bar{N}$ line segments over the intervals

$$[M_{\ell+1}, M_\ell], \quad \ell \in \{1, \ldots, \bar{N} - 1\} \quad (39a)$$

$$[M_1, M_0), \quad (39b)$$

where the last interval is half-open, see (5).

We next upper bound $R_{YMA}(K, N, M)$ by a function $\overline{R}_{YMA}(K, N, M)$ that is piecewise-linear over the same intervals (39). Specifically, for every $\ell \in \mathcal{N}$, define for $M \in [M_\ell, M_{\ell-1})$:

$$R_{YMA}(K, N, M) := \frac{M - M_\ell}{M_{\ell-1} - M_\ell} \cdot R_{YMA}(K, N, M_\ell) + \frac{M_{\ell-1} - M}{M_{\ell-1} - M_\ell} \cdot R_{YMA}(K, N, M_{\ell-1}).$$

Notice that

$$R_{YMA}(K, N, M_\ell) = R_{YMA}(K, N, M_\ell), \quad \forall \ell \in \{1, \ldots, \bar{N}\},$$

whereas for general $M \in [0, N)$:

$$R_{YMA}(K, N, M) \geq R_{YMA}(K, N, M_\ell),$$ (40)

because $R_{YMA}(K, N, M)$ is convex.

Plugging (40) into (36), we obtain:

$$\xi(K, N, M) \leq \frac{R_{YMA}(K, N, M)}{R_{\text{worst}}(K, N, M)} =: \Xi(K, N, M).$$ (41)

Now, since the upper bound $\Xi(K, N, M)$ is continuous and bounded in $M \in [0, N)$, and because it is quasiconvex [59] in $M$ over each of the $\bar{N}$ intervals (39), the maximum of $\Xi(K, N, M)$ over each interval is attained at the boundary. Furthermore, since $\overline{R}_{YMA}(K, N, M_0) = R_{\text{worst}}(K, N, M_0) = 0$, for $M \in [M_1, M_0)$:

$$\Xi(K, N, M) = \frac{M_1 - M_\ell}{M_{\bar{N}} - M_\ell} \cdot \overline{R}_{YMA}(K, N, M_1) + \frac{M_{\bar{N}} - M}{M_{\bar{N}} - M_1} \cdot R_{\text{worst}}(K, N, M_1)$$

$$= \overline{R}_{YMA}(K, N, M_1) = \Xi(K, N, M_1).$$ (42)

Thus, $\Xi(K, N, M)$ is constant over the last interval $[M_1, M_0)$ and we obtain that

$$\max_{M \in [0, N)} \xi(K, N, M) \leq \max_{\ell \in \mathcal{N}} \Xi(K, N, M_\ell).$$ (43)

A linear-fractional function is always quasiconvex.
Irrespective of \( K, N \in \mathbb{Z}^+ \), we have:

\[
\Xi(K, N, M) = \frac{\tilde{N}}{N} = 1.
\] (44)

When \( \tilde{N} = 1 \) (i.e., only one file or only one user), Inequalities (43) and (44) imply that the gap \( \xi(K, N, M) = 1 \) for all \( M \in [0, N) \), and hence our lower bound is exact.

We therefore assume in the following that \( \tilde{N} \geq 2 \). For \( \ell \in \{1, \ldots, \tilde{N} - 1\} \), we have

\[
\Xi(K, N, M_\ell) = \frac{\sum_{j=1}^{\ell} \left( 1 - \left( \frac{1}{N} \right)^{\frac{N-\ell}{\ell+1}} \right)}{\ell - (N - \ell) \sum_{j=1}^{\ell} \frac{1}{N-j+1}} \leq \frac{\ell}{\ell - (N - \ell) \sum_{j=1}^{\ell} \frac{1}{N-j+1}}
\]

\[
a = \frac{\ell}{\ell - (N - \ell) \sum_{j=1}^{\ell} \frac{1}{N-j+1}} \leq \phi(a, \ell).
\] (45)

Note that since \( \ell \in \{1, \ldots, \tilde{N} - 1\} \),

\[
a \in \left[\frac{1}{N}, 1\right).
\]

Therefore,

\[
\max_{K \in \mathbb{Z}^+} \max_{N \in \mathbb{Z}^+} \max_{M \in [0, N)} \xi(K, N, M) \leq \max_{\ell \in \mathbb{Z}^+} \max_{a \in (0, 1)} \phi(a, \ell),
\]

which concludes the proof.

**B. Proof of Inequality (7)**

We have a closer look at the denominator of the function \( \phi(a, \ell) \).

Notice that \( \frac{1}{n} \leq \int_{n-1}^{n} \frac{dt}{t} \) for all \( n \geq 2 \). Therefore,

\[
\sum_{j=0}^{\ell-1} \frac{1}{\ell-a_j} = \sum_{j=0}^{\ell-1} \frac{1}{\ell(1-a) + a(\ell-j)} = \sum_{i=0}^{\ell} \frac{1}{\ell(1-a) + ai} = \frac{1}{\ell(1-a) + a} + a \sum_{i=2}^{\ell} \frac{1}{\ell(1-a)/a + i}
\]
\[
\begin{align*}
\frac{1}{\ell(1-a) + a} + a \sum_{i=2}^{\ell} \int_{(1-a)/a+i-1}^{(1-a)/a+i} \frac{1}{t} dt \\
= \frac{1}{\ell(1-a) + a} + \frac{1}{a} \int_{(1-a)/a+1}^{(1-a)/a+\ell} \frac{1}{t} dt \\
= \frac{1}{\ell(1-a) + a} + \frac{1}{a} \ln \left( \frac{\ell}{\ell(1-a) + a} \right).
\end{align*}
\] (46)

We use (46) to upper bound the function \( \phi(a, \ell) \):

\[
\phi(a, \ell) = \frac{a(\ell+1)}{(1-a)\ell} \left( 1 - \left( \frac{\ell+a}{\ell+1} \right)^{\ell/a} \right)
\]

\[
\leq \frac{a(\ell+1)}{(1-a)\ell} \left( 1 - \left( \frac{\ell+a}{\ell+1} \right)^{\ell/a} \right)
\]

\[
= \frac{a(1+b)}{1-a} \left( 1 - \left( \frac{1+ab}{1+b} \right)^{2\ell} \right) = \psi(a, b).
\]

Noting also that if \( \ell > 10^4 \), then \( b < 10^{-4} \), this concludes the proof of (7).

VI. PROOF OF THE GAP-RESULT IN THEOREM 5

We wish to uniformly bound the gap

\[
\theta(K, N, M) := \frac{R^*_\text{avg}(M)}{R^\text{low}_\text{avg}(M)},
\] (47)

irrespective of \( K, N \geq 1 \) and \( M \in [0, N) \).

Since \( R_Y^{\text{MA}}(K, N, M) \) upper bounds the rate-memory tradeoff for the worst case, it must also upper bound the rate-memory tradeoff for the average case. Thus, for any number of users \( K \) and files \( N \):

\[
R^*_\text{avg}(M) \leq R_Y^{\text{MA}}(K, N, M), \quad M \in [0, N). \] (48)

We thus have

\[
\theta(K, N, M) \leq \frac{R_Y^{\text{MA}}(K, N, M)}{R^\text{low}_\text{avg}(M)} \leq \frac{R_Y^{\text{MA}}(K, N, M)}{R^\text{avg}(K, N, M)},
\] (49)

where we defined

\[
R^\text{avg}(K, N, M) := \max_{\ell \in N} \sum_{k=1}^{\ell} \left[ \left( 1 - \frac{1}{N} \right)^{k-1} - \frac{k}{N} M \right],
\]

and where the second inequality holds because for all \( K, N, M \):

\[
R^\text{low}_\text{avg}(M) \geq \max_{\ell \in N} \left[ \left( 1 - \left( 1 - \frac{1}{N} \right)^{\ell} \right) N - \frac{\ell(\ell+1)}{2N} M \right]
\]

\[
= \max_{\ell \in N} \sum_{k=1}^{\ell} \left[ \left( 1 - \frac{1}{N} \right)^{k-1} - \frac{k}{N} M \right].
\]
Define
\[ \tilde{M}_\ell := \begin{cases} \frac{N}{e^{1/\ell}} \left( 1 - \frac{1}{N} \right)^\ell & \text{if } \ell \in \{0, 1, \ldots, \tilde{N} - 1\}, \\ 0 & \text{if } \ell = \tilde{N}. \end{cases} \]

Note that 0 = \tilde{M}_{\tilde{N}} < \tilde{M}_{\tilde{N} - 1} < \cdots < \tilde{M}_0 = N. The function \( \bar{R}_{\text{avg}}(K, N, M) \) is piecewise-linear with \( \tilde{N} \) line segments over the intervals
\[ \left[ \tilde{M}_{\ell + 1}, \tilde{M}_\ell \right], \quad \ell \in \{1, \ldots, \tilde{N} - 1\} \] \[ \left[ \tilde{M}_1, \tilde{M}_0 \right), \]
where the last interval is half-open, see (5).

We next upper bound \( R_{\text{YMA}}(K, N, M) \) by a function \( \bar{R}_{\text{YMA}}(K, N, M) \) that is piecewise-linear over the same intervals (50). Specifically, for every \( \ell \in \tilde{N} \), define for \( M \in [\tilde{M}_\ell, \tilde{M}_{\ell - 1}) \):
\[ \bar{R}_{\text{YMA}}(K, N, M) := \frac{M - \tilde{M}_\ell}{M_{\ell - 1} - M_\ell} \cdot R_{\text{YMA}}(K, N, \tilde{M}_\ell) \]
\[ + \frac{\tilde{M}_{\ell - 1} - M}{M_{\ell - 1} - M_\ell} \cdot R_{\text{YMA}}(K, N, \tilde{M}_{\ell - 1}). \]

Notice that
\[ \bar{R}_{\text{YMA}}(K, N, \tilde{M}_\ell) = R_{\text{YMA}}(K, N, \tilde{M}_\ell), \quad \forall \ell \in \{1, \ldots, \tilde{N}\}, \]
whereas for general \( M \in [0, N] \):
\[ \bar{R}_{\text{YMA}}(K, N, M) \geq R_{\text{YMA}}(K, N, M), \] because \( R_{\text{YMA}}(K, N, M) \) is convex.

Plugging (51) into (49), we obtain:
\[ \theta(K, N, M) \leq \frac{R_{\text{YMA}}(K, N, M)}{\bar{R}_{\text{avg}}(K, N, M)} =: \Theta(K, N, M). \]

Following similar arguments as in the proof of Theorem 2, we have
\[ \max_{M \in [0, N]} \theta(K, N, M) \leq \max_{\ell \in \tilde{N}} \Theta(K, N, \tilde{M}_\ell). \] (53)

Irrespective of \( K, N \in \mathbb{Z}^+ \), we have:
\[
\Theta(K, N, \tilde{M}_{\tilde{N}}) = \frac{\tilde{N}}{N(1 - (1 - 1/N)^{N})} \quad \quad \text{for } \ell \in \{1, \ldots, \tilde{N} - 1\},
\]
\[ \leq \frac{x}{1 - e^{-x}} \bigg|_{x = \tilde{N}/N} \] (a)
\[ \leq \frac{1}{1 - e^{-1}}, \] (b)

where (a) follows since \((1 - 1/\zeta)^\ell \leq e^{-1}\) for all \( \zeta > 1 \) and (b) follows since \( x \mapsto \frac{x}{1 - e^{-x}} \) is an increasing function. This implies that when \( \tilde{N} = 1 \) (i.e., one file or one user), then \( \Theta(K, N, M) \leq \frac{1}{1 - e^{-1}} \leq 1.582 \) for all \( M \in [0, N] \).

In the following, we assume that \( \tilde{N} \geq 2 \). As for \( \ell \in \{1, \ldots, \tilde{N} - 1\} \), we have
\[
\Theta(K, N, \tilde{M}_\ell) = \frac{N - \frac{\tilde{N}}{e^{1/\ell}} (1 - \frac{1}{N})^\ell}{N - (N + \frac{\ell}{2}) (1 - \frac{1}{N})^\ell} \left( 1 - \left( 1 - \frac{1}{N} \frac{\tilde{N}}{e^{1/\ell}} (1 - \frac{1}{N})^\ell \right)^{\tilde{N}} \right).
\]
which concludes the proof.

Also, it holds that

\[
\frac{(\ell + 1 - N (1 - 1/N)^{\ell})}{(1 - 1/N)^{\ell}(1 - (1 + \ell/2N)(1 - 1/N)^{\ell})} \leq \frac{(\ell + 1 - N (1 - 1/N)^{\ell})}{(1 - 1/N)^{\ell}(1 - (1 + \ell/2N)(1 - 1/N)^{\ell})} \leq \left(\frac{u + v - v (1 - v)u/v}{1 - (1 - v/u + v (1 - v)u/v)^{1/u}}\right)
\]

\[
= \frac{(1 - v)^{u/v}(1 - (1 + u/2) (1 - v)^{u/v})}{(1 - v)^{u/v}(1 - (1 + u/2) (1 - v)^{u/v})} =: \eta(u, v),
\]

where (a) follows by a change of variable \( u = \ell/N \) and \( v = 1/N \). Note that since \( \ell \in \{1, \ldots, N - 1\} \) and assuming \( N \geq 2 \),

\[
u \in \left[\frac{1}{N}, \frac{N - 1}{N}\right]\text{ and } v \in (0, 1/2].
\]

Also, it holds that \( \frac{1}{1-e^{-x}} < \max_{u \in [0,1]} \max_{v \in (0,1/2]} \eta(u, v) \). Therefore,

\[
\max_{K \in \mathbb{Z}^+} \max_{N \in \mathbb{Z}^+} \max_{M \in \{0, N\}} \theta(K, N, M) \leq \max_{u \in [0,1]} \max_{v \in (0,1/2]} \eta(u, v),
\]

which concludes the proof.

REFERENCES


